### Title

Self-Attention Convolutional Neural Network for Improved MR Image Reconstruction.




### Authors

Wu, Yan
Ma, Yajun
Liu, Jing
et al.



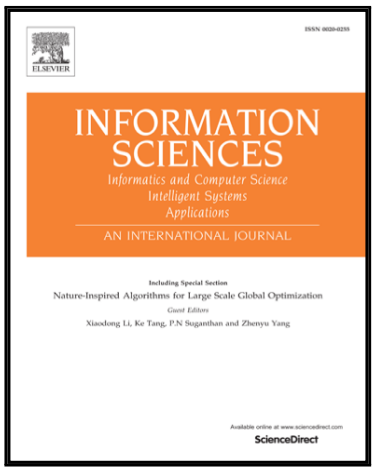
### Publication Date

2019-07-01

### DOI

10.1016/j.ins.2019.03.080

Peer reviewed




# Self-Attention Convolutional Neural Network for Improved MR Image Reconstruction

Yan Wu , Yajun Ma , Jing Liu , Jiang Du , Lei Xing

PII: S0020-0255(19)30291-9
DOI: https://doi.org/10.1016/j.ins.2019.03.080
Reference: INS 14420

To appear in: *Information Sciences*

Received date: 17 December 2018
Revised date: 29 March 2019
Accepted date: 31 March 2019

Please cite this article as: Yan Wu , Yajun Ma , Jing Liu , Jiang Du , Lei Xing , Self-Attention Convolutional Neural Network for Improved MR Image Reconstruction, *Information Sciences* (2019), doi: https://doi.org/10.1016/j.ins.2019.03.080

# Self-Attention Convolutional Neural Network for Improved MR Image Reconstruction

**Yan Wu[1], Yajun Ma[2], Jing Liu[3], Jiang Du[2], Lei Xing[1*]**

1. Radiation Oncology Department, Stanford University. 875 Blake Wilbur Drive G204, Stanford, California 94305
2. Radiology Department, University of California San Diego. 9500 Gilman Driven #0888, La Jolla, California 92093
3. Radiology Department, University of San Francisco. 185 Berry St, San Francisco, California 94107

* corresponding author: Lei Xing

lei@stanford.edu,

875 Blake Wilbur Drive Room G204, Stanford, California, USA, 94305

## Abstract

MRI is an advanced imaging modality with the unfortunate disadvantage of long data acquisition time. To accelerate MR image acquisition while maintaining high image quality, extensive investigations have been conducted on image reconstruction of sparsely sampled MRI. Recently, deep convolutional neural networks have achieved promising results, yet the local receptive field in convolution neural network raises concerns regarding signal synthesis and artifact compensation. In this study, we proposed a deep learning-based reconstruction framework to provide improved image fidelity for accelerated MRI. We integrated the self-attention mechanism, which captured long-range dependencies across image regions, into a volumetric hierarchical deep residual convolutional neural network. Basically, a self-attention module was integrated to every convolutional layer, where signal at a position was calculated as a weighted sum of the features at all positions. Furthermore, relatively dense shortcut connections were employed, and data consistency was enforced. The proposed network,

referred to as SAT-Net, was applied on cartilage MRI acquired using an ultrashort TE sequence and retrospectively undersampled in a pseudo-random Cartesian pattern. The network was trained using 336 three dimensional images (each containing 32 slices) and tested with 24 images that yielded improved outcome. The framework is generic and can be extended to various applications.

**Introduction**

Magnetic Resonance Imaging (MRI) is an advanced imaging modality that provides superior soft tissue contrast. The primary disadvantage of MRI, however, is the long data acquisition time. To accelerate image acquisition, various sparse sampling (i.e. sampling fewer data points in the sensor domain or k-space in order to reduce scan time) schemes were proposed (Peters, 2000; Pike,1994); however, these techniques introduced image blurring and undersampling artifacts. To improve image fidelity in accelerated data acquisition, advanced image reconstruction techniques were developed. Specifically, compressed sensing (CS) became a breakthrough method in the past decade (Lustig, 2007). In some variants of compressed sensing, *a priori* information was actively exploited and incorporated into image reconstruction (Bilgic, 2001; Vaswani, 2010).

Non-patient specific *a priori* information can be acquired via deep learning (LeCun, 2015) and utilized for image reconstruction. In recent years, deep learning has led to a flood of breakthroughs in image processing and begun to change the landscape of medical physics (Xing, 2018). Particularly, convolutional neural networks (CNNs) and CNN-based Generative Adversarial Networks (GAN) have been influential in medical imaging (Krizhevsky, 2012; Goodfellow, 2014). In some pilot studies on image reconstruction of sparsely sampled MRI, *a priori* information acquired through deep learning was incorporated into the framework of compressed sensing, either as the initial image or as the optimal parameters defined in the model. Hammernik *et al*. employed a convolutional neural network to find the optimal parameters specified at different stages of a variational network, which mimicked different iterations in compressed sensing processing (Hammernik. 2018). Yang *et al.* used a non-convolutional neural

network to search for the best parameters defined in the ADMM (Alternating Direction Method of Multipliers) and compressed sensing model (Yang, 2017). Wang *et al.* utilized a convolutional neural network to initialize the compressed sensing model in a two-phase reconstruction or was integrated into the compressed sensing framework as an additional regularization term (Wang, 2016). Yang *et al.* incorporated the conditional GAN loss into the compressed sensing framework as a regularization term (Yang, 2018).

Alternatively, deep neural networks have been proposed to provide an end-to-end mapping from sparsely sampled k-space data or images to fully sampled images. Zhu *et al.* constructed an AUTOMAP (automated transform by manifold approximation) framework, which was composed of a few fully connected layers and some convolutional layers (Zhu, 2018). Schlemper *et al.* proposed a deep network architecture, which was formed by a sequence of convolutional neural networks with data consistency enforced on the output of every network (Schlemper, 2018). Ke *et al.* developed a cascaded residual dense network for cardiac MR image reconstruction, which included both k-space prediction networks and spatial-domain residual dense networks to enable cross-domain learning (Ke, 2019). Lyu *et al.* established a deep neural network for quantitative relaxation parametric mapping, where different k-space lines contained various signal contrast (Lyu, 2018). Liu *et al.* built a deep neural network for super-resolution within 2D slices, which adopted deconvolution operation instead of bicubic interpolation as the preprocessing step (Liu, 2018). Chaudhari *et al.* employed a three-dimensional patched-based convolutional neural network for super-resolution in the slice encoding direction (Chaudhari, 2018).

While convolutional neural networks achieved high performance because of the effective extraction of local image features, the convolution operator has a limited range of influence or local receptive field. This can be problematic for synthesizing signal from a wide range of inputs. The receptive field can be enlarged by increasing the depth of a convolutional neural network or employing a hierarchical network architecture. However, a deep stack of convolutional operations is not only computationally prohibitive for volumetric high-resolution images, but also introduces difficulties in

optimization when long-range dependencies are progressively propagated across multiple layers (Hochreiter, 1997).

Furthermore, it would be preferable to efficiently use the available information. For example, in a deep learning-based segmentation approach, supervoxel-based partition was applied to identify critical regions close to the boundary, such that more computational efforts could be made within the critical regions (Qin, 2018). Alternatively, in various models that incorporated the self-attention mechanism (Parikh, 2016; Luong, 2015), signal at a position was obtained by attending to all positions in the same image (rather than near neighbors) with weights determined by the similarity between voxels. In this way, distant voxels were allowed to make direct contributions, facilitating long-range dependencies.

In fact, it has been a trend to integrate the self-attention mechanism, which captures long-range dependencies, into deep neural networks. An innovative network architecture called Transformer (Vaswani, 2017) was proposed for machine translation applications, where a stack of building blocks was employed, each composed of one or more self-attention layers and a fully connected layer. The Transformer model was tailored to the Image Transformer model for image synthesis tasks, where 'local' self-attention maps were derived from small image patches to relieve the heavy computation load caused by a large number of voxels in an image (Parmar, 2018). Meanwhile, the non-local neural network architecture was developed for video classification, which viewed the self-attention mechanism as a special case of non-local filtering operations that captured global dependencies (Wang, 2017). Along the same direction, the self-attention GAN (Zhang, 2018) was constructed for the generation of natural images, where the self-attention mechanism was incorporated into both generator and discriminator of a convolutional GAN. Attention was used in other applications, such as similarity learning (Gao, 2018) and hand gesture recognition (Li 2018).

In this study, we proposed a self-attention convolutional neural network for MRI reconstruction, which introduced the self-attention mechanism into a deep

convolutional neural network. At every convolution layer, self-attention maps were derived from feature maps, where the attention value at a position was calculated as a weighted sum of the features at all positions. Hence, the self-attention module was able to model long-range dependencies across image regions and was complementary to local convolution operators. We implemented the self-attention mechanism on top of a high performance convolutional neural network, T-Net, which was a volumetric hierarchical deep residual network with established global and local shortcut connections. Data consistency was further enforced in k-space, where the network predictions were replaced by original measurements at the data points that were actually sampled. This self-attention (SA) T-Net was referred to as SAT-Net.

While the SAT-Net provides a generic framework for MR image reconstruction, in this study, it was mainly applied to cartilage MRI that was acquired using an ultrashort TE (UTE) sequence and retrospectively undersampled using a pseudo-random Cartesian sampling scheme. Using the SAT-Net, image quality was well maintained when a high acceleration factor of 6 was achieved. Particularly, the self-attention mechanism demonstrated its ability to improve the reconstruction outcome.

**Method**

In this study, a self-attention convolutional neural network, SAT-Net, was developed for the reconstruction of sparsely sampled MRI. With Institutional Review Board approval and HIPAA compliance, 360 three dimensional cartilage images were acquired using a special ultra-short TE (UTE) sequence and retrospectively undersampled using a pseudo-random Cartesian sampling. SAT-Net was used to provide an end-to-end mapping from sparsely sampled images to their fully sampled correspondence with data consistency further enforced in k-space. The workflow is illustrated in Figure 1.

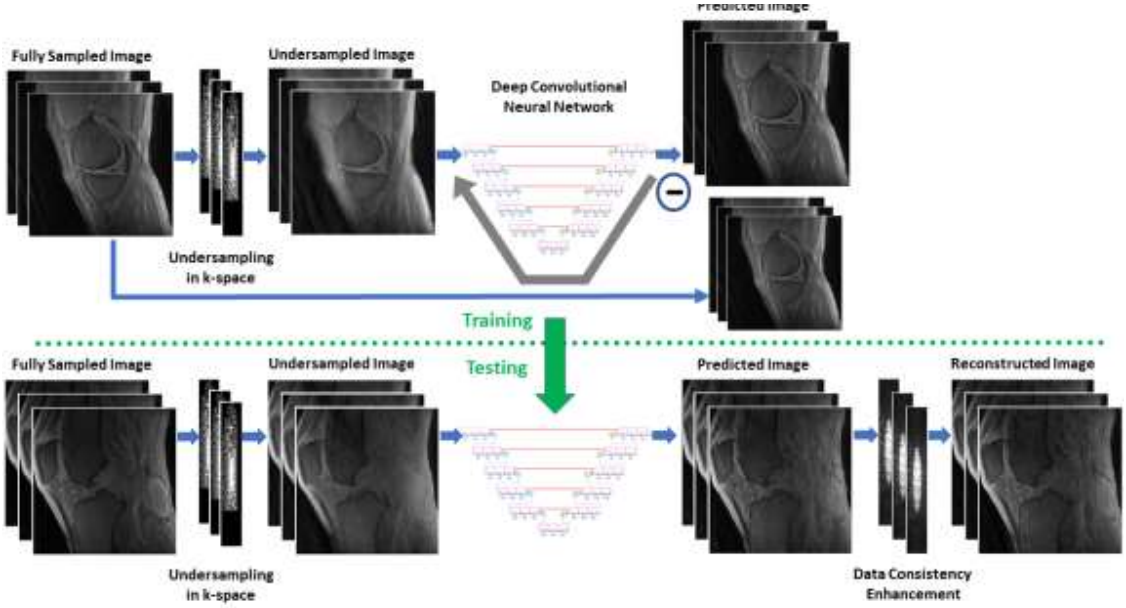



**Figure 1**. The workflow of employing the self-attention convolutional neural network (SAT-Net) for MR reconstruction. Fully sampled 3D images were retrospectively undersampled in k-space and transformed back to the image domain. In the training procedure, a convolutional neural network was established, in which loss was back-propagated and used to update model parameters iteratively. In testing, retrospectively undersampled images were passed through the well-trained network, and subsequent data consistency enforcement was performed to form the final reconstruction result.

### *Image Acquisition and Sparse Sampling*

Three hundred and sixty 3D cartilage images were acquired at the University of California San Diego (Ma, 2018). The data were acquired on a 3T scanner (GE Healthcare, Waukesha, WI) using an adiabatic inversion recovery spin-lock prepared UTE sequence with different numbers of IR spin-lock pulses (2, 4, 6, 8, 12, and 16). Other imaging parameters were as follows: echo time of 32 μs, repetition time of 500 ms, flip angle of 10°, resolution of 256x256x32, voxel size of 0.586x0.586x3 mm, and a scan time of 2.7 min per data set.

Given fully sampled images, a pseudo-random variable-density Cartesian acquisition, CIRcular Cartesian UnderSampling (CIRCUS) was simulated (Liu, 2014), as illustrated in Figure 2. Sparse sampling was performed on the ky-kz plane with an acceleration factor of 6 achieved. The undersampled k-space data were transformed back to the image

domain using the 3D inverse Fourier Transform, and taken as the input to the neural network.

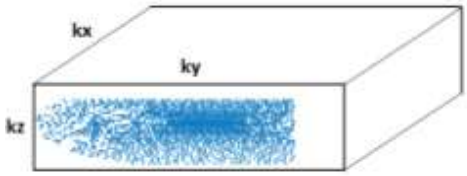



**Figure 2**. The pseudo-random variable-density Cartesian sampling pattern (CIRCUS). Undersampling was performed on the ky-kz plane, with an acceleration factor of 6 achieved.

***Reconstruction Using the Self-Attention Convolutional Neural Network***

A self-attention convolutional neural network (SAT-Net) was developed for MRI reconstruction, which introduced the self-attention mechanism (SA) into a volumetric hierarchical deep residual convolutional neural network.

Basically, a deep convolutional neural network was employed to provide a data-driven end-to-end mapping from a sparsely sampled MR image to its corresponding fully sampled image. Throughout the network, volumetric processing was adopted to fully exploit 3D spatial continuity. While the utilization of global information was critical for signal synthesis and artifact compensation, a larger receptive field would improve outcomes significantly.

To enlarge the receptive field, a hierarchical deep neural network was constructed. The network was composed of an encoder (contracting path with a decreased resolution of feature maps) and a decoder (expanding path with an increased resolution of feature maps), both having multiple levels. At each level, the resolution of feature maps was kept the same. From one level to the next along the contracting path, feature maps were down-sampled, and the number of feature maps (convolution kernels) was doubled as indicated. In the expanding path, the change in image resolution and number of feature maps was reversed. Both down-sampling and up-sampling were accomplished using convolution operations (with 2x2x2 kernels) instead of conventional

pooling, as suggested by (Springenberg, 2014). The network architecture is shown in Figure 3.

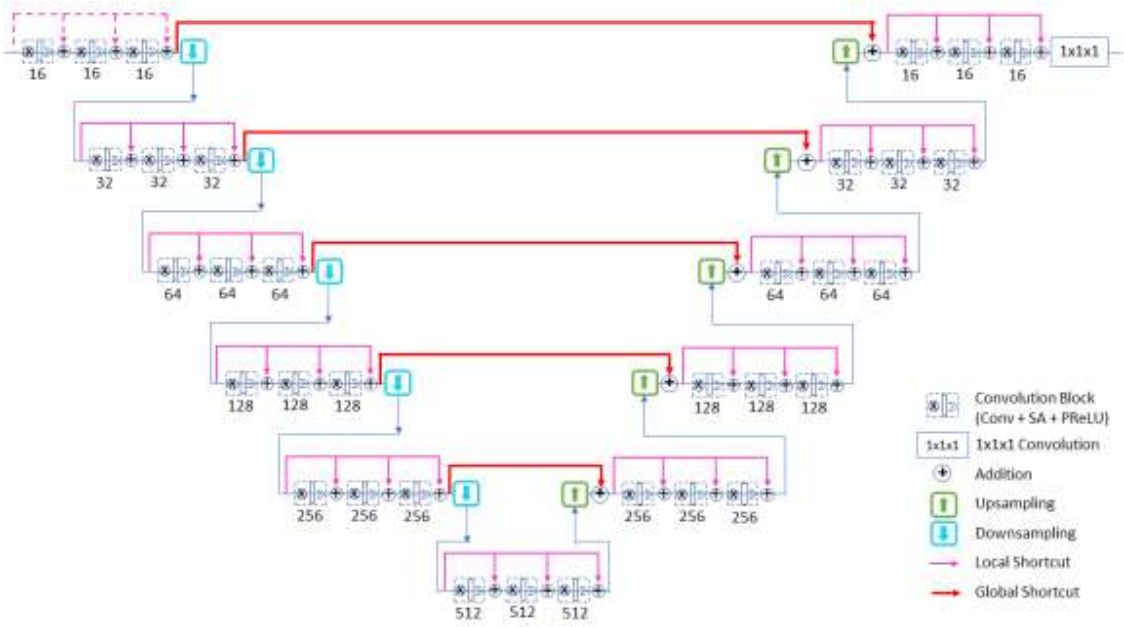



**Figure 3**. The architecture of the volumetric hierarchical deep residual convolutional neural network (SAT-Net). It is composed of a contracting path (on the left) and a subsequent expanding path (on the right), along which the resolution of feature maps first shrinks and then expands. Global shortcut connections are established between the corresponding levels of the two paths, whereas local shortcut connections are established within the same level of a single path.

In addition, 'global' shortcuts connected the corresponding levels of the encoder and the decoder to compensate for details lost in down-sampling, whereas 'local' shortcut connections were established within the same level of a single path to facilitate residual learning (He, 2016). Moreover, relatively dense local shortcut connections were formed in this SAT-Net by forwarding the input of a hierarchical level to all the subsequent convolutional blocks at the same level, as illustrated in Figure 4. In contrast, U-Net (Ronneberger, 2015) had no local shortcut connections, whereas V-Net (Milletari, 2016) had simple local shortcut connections. The relatively dense design was inspired by Dense Net [34, 35] (Huang, 2017; Lodhi, 2019) and Deep Recursive Residual Network (Tai, 2017), as illustrated in Figure 5, where the former demonstrated the effect of

dense shortcut connections on the improvement of performance, and the latter proposed an alternative connection pattern that was more computationally efficient. We constructed the relatively dense local shortcut connection pattern to reach a good balance between the network performance and memory consumption.

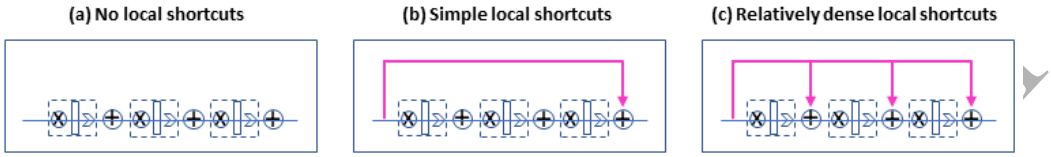



**Figure 4**. Comparison of our local shortcut connection scheme with the ones adopted in U-Net and V-Net. (a) No local shortcuts, as in U-Net, (b) simple local shortcuts (forwarding the input of a hierarchical level to the output at the same level), as in V-Net, and (c) relatively dense local shortcuts (forwarding the input of a hierarchical level to all the subsequent convolutional blocks at the same level), as we proposed in SAT-Net.

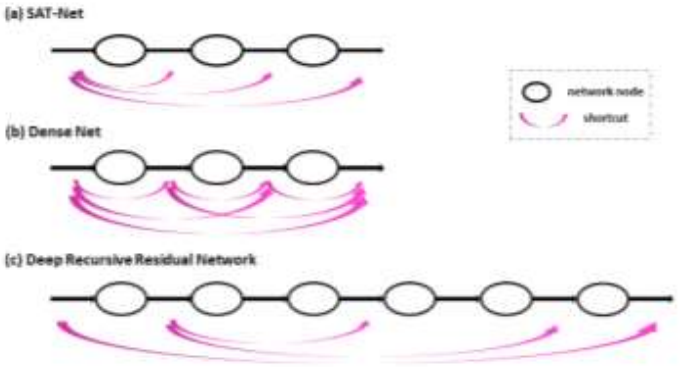



**Figure 5.** Several comparatively dense shortcut connection schemes that inspired our design. (a) Our SAT-Net, in which the input of a network level was forwarded to all the subsequent convolutional blocks at the same level, (b) Dense Net, in which the output of every convolutional block was forwarded to all the subsequent blocks, and (c) Deep Recursive Residual Network, which had shortcut connections with various ranges of influence and the origins of the shortcuts were close to the input of the network level.

At every level in the hierarchical network, there were three convolutional blocks. Each convolutional block was composed of a convolutional layer that extracted image features and of an activation layer that provided nonlinearity.

Based on this T-Net, the self-attention mechanism was incorporated into every convolutional block, as shown in Figure 3. In this way, global information that was

spread in widely separated spatial regions of feature maps could be efficiently utilized. In the resultant SAT-Net, a convolutional block was composed of three layers, a convolution layer that captured local information, a novel self-attention layer that supplied non-local information, and a nonlinear activation layer which was implemented by parametric ReLU function (He, 2015). Figure 6 compared the convolution block of a self-attention convolutional neural network with that of a traditional network.

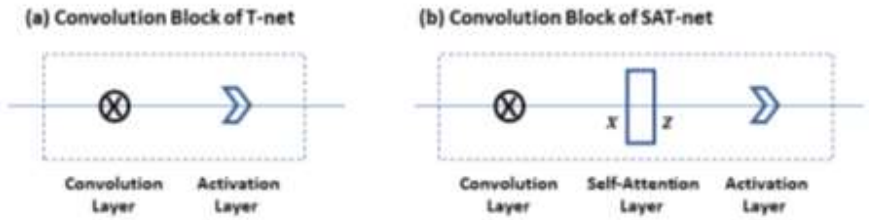



**Figure 6**. Comparison of the convolution block of a self-attention convolutional neural network with that of a traditional network. (a) Traditional convolution block without the self-attention mechanism incorporated, as used in the T-Net, and (b) novel convolution block with the self-attention mechanism incorporated, as used in the SAT-Net, which provides long-range dependencies.

***Implementation of Self-Attention Modules***

The self-attention maps within each convolution block were derived for feature maps extracted at the preceding convolution layer. The attention value at a position was obtained by attending to all positions in the same image with different weights, formulated as

$$Y_i = \frac{1}{C(X_i)} \sum_j s(X_i, X_j) h(X_j)$$

(1)

where $i$ was a position at which the response was computed, $j$ enumerated all positions in the same image, function $s$ calculated a scalar that revealed the relevance between the signal intensity at the current position $i$ and that at any position $j$, and function $h$ computed a representation of the input signal at any position $j$. That meant only signal at relevant positions contributed directly to the signal at the current position, and the

contribution was determined by both the relevance and the signal intensity at the distant positions. The response was subsequently normalized by a factor $C(X_i)$, where

$$C(X_i) = \sum_j s(X_i, X_j) \quad .$$

(2)

The functions $s$ and $h$ were chosen with some degree of flexibility, since networks that modeled long-range dependencies were found to be insensitive to the choice (Wang, 2017). In this work, $h$ was defined as a linear function for simplicity

$$h(X_j) = W_h X_j \quad (3)$$

and $s$ was defined as an embedded Gaussian function because Gaussian function was a natural choice to quantify the similarity (relevance) between $X_i$ and $X_j$:

$$s(X_i, X_j) = e^{\emptyset(X_i)^T \theta(X_j)} = e^{(W_f X_i)^T (W_g X_j)} \quad .$$

(4)

Here, $W_f$, $W_g$, and $W_h$ were weight matrices implemented as 1×1×1 convolution, which were learned in the training of the network.

In addition, residual learning was employed in the attention layer, as given by

$$Z_i = X_i + \alpha Y_i. \quad (5)$$

Therefore, the output of the self-attention layer ($Z_i$) was composed of two components: one was the feature maps from the previous convolution layer that captured local information ($X_i$) and the other was the self-attention maps that provided non-local information ($Y_i$). A scale parameter $\alpha$ balanced the contributions from local and non-local sources in the response, which was learned during the training. Initially, $\alpha$ was set to 0, and the SAT-Net had a learning behavior similar to that of a convolutional network. As $\alpha$ was increased during optimization, the self-attention layers gradually took effect, accomplishing a smooth transition to the distinguished self-attention convolutional neural network. In this way, the attention layer was seamlessly integrated, enabling

efficient utilization of information from widely separated spatial regions. The structure of a self-attention layer is shown in Figure 7.

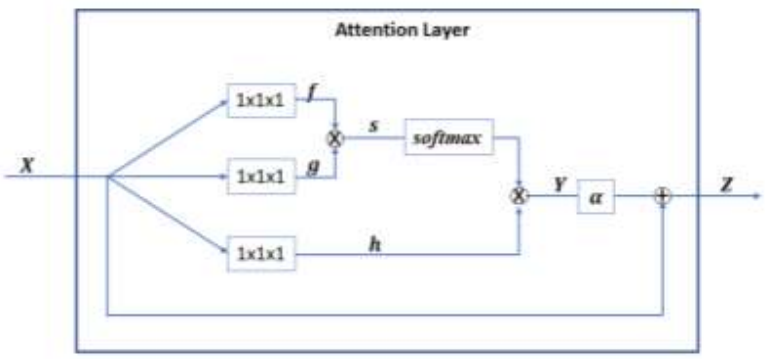



**Figure 7**. Structure of a self-attention layer. The output of a self-attention layer is composed of two components: one is the feature maps from the previous convolution layer that capture local information and the other is the self-attention maps that provide non-local information. A scale parameter $\alpha$ balances the contributions from local and non-local sources in the response. In the attention maps, function $s$ computes a scalar that reveals the relevance between the signal intensity at the current position $i$ and at any position $j$, and function $h$ computes a representation of the input signal at any position $j$.

***Data Consistency Enforcement***

While the output of the self-attention convolutional neural network provided reasonable estimates for k-space coefficients at all data points, it would be more accurate to replace the network predictions with original measurements at the data points that were actually sampled. In the framework of deep learning, this was incorporated into the Root Mean Squared Error (RMSE) loss function (de Jesús Rubio, 2018; Meng, 2018; de Jesús Rubio, 2009; Zhang 2018; de Jesús Rubio, 2017; Jiang, 2018),

$$RMSE = \sqrt{\left(\sum_{i=1}^{n} \|z_i - x_i\|^2\right)}, \quad (6)$$

where $x_i$ was the fully sampled image and $z_i$ was the predicted image with data consistency enforced. $z_i$ was formulated by

$$z_i = IFT\{d(k_i, \widehat{k_i})\} \quad (7)$$

in which $k_i$ was the measured k-space data, $\widehat{k_i}$ was the estimated k-space data obtained from the network predictions at the current iteration, and $h$ was the data consistency

enforcement function. The estimated k-space data $\hat{k}$ was the Fourier transform of the network output, as given by

$$\hat{k} = FT(f_{cnn}(x_{ZF}|\theta)) \quad (8)$$

where $f_{cnn}$ was the forward mapping of the convolutional neural network that took undersampled zero-filled image $x_{ZF}$ as the input. The data consistency enforcement function $d$ was defined as

$$d(k_i, \widehat{k_i}) = I_{\{0\}}(k_i) * \widehat{k_i} + (1 - I_{\{0\}}(k_i)) * k_i \quad (9)$$

with $I$ representing the indicator function. More intuitively, $d$ could be expressed as

$$d(k_i, \widehat{k_i}) = \begin{cases} k_i \ \ if\ k_i \neq 0 \\ \widehat{k_i}\ if\ k_i = 0 \end{cases} .$$

(10)

The indicator function was not continuous, but it did not affect the performance of the convolutional neural network [15]. The self-attention convolutional neural network was implemented on a tensor-flow (Abadi , 2016) based AI platform, NiftyNet (Gibson, 2018).

***Training and Testing of the Self-Attention Convolutional Neural Network***

The SAT-Net was trained to learn the optimal network parameters using 336 three dimensional images (each containing thirty-two slices). Data augmentation was performed, including translation, rotation, and flipping, to increase the quantity of training data. The network parameters were initialized using the He method (He, 2015) and updated using the Adam algorithm (Kingma, 2014) with an adaptive learning rate (starting from 0.001, $\beta_1$ of 0.9, $\beta_2$ of 0.999, and $\epsilon$ of $10^{-8}$).

Given the trained neural network, 24 three dimensional images from different subjects were tested. The quality of reconstructed images was evaluated both qualitatively and quantitatively. We measured two quantitative metrics. One was the Structural Similarity Index Measure (SSIM), which measured the perceptual difference between two similar images, and the other was the Peak Signal-to-Noise Ratio (PSNR). SSIM was defined as

$$\mathrm{SSIM}(x,y)=\frac{(2\mu_x\mu_y+C_1)(2\sigma_x\sigma_y+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$$

(11)

where$\mu_x$, $\mu_y$, $\sigma_x$, and $\sigma_y$corresponded to the mean and standard deviation of signal intensity in the reconstructed image and the ground truth, $C_1=(k_1L)^2$, $C_2=(k_2L)^2$, $k_1=0.01$, $k_2=0.03$, and $L$ is the dynamic range of the pixel values. PSNR was defined as

$$\mathrm{PSNR}=10\times\log_{10}\left(\frac{\left(2^n-1\right)^2}{\mathrm{MSE}}\right)$$

(12)

where MSE is the mean squared error, defined as

$$MSE=\sum_{i=1}^{n}\|z_i-x_i\|^2 \quad .$$

(13)

**Results**

Trained with 336 three dimensional images and tested with 24 images, the SAT-Net provided improved image fidelity, with a high acceleration factor of 6 achieved in cartilage MRI. In the pseudo-random Cartesian acquisition, high frequency information lost due to sparse sampling was substantially recovered. Particularly with the self-attention mechanism incorporated into the convolutional neural network, the quality of reconstructed images was significantly improved.

Figure 8 compared sparsely sampled MR images reconstructed with and without the self-attention mechanism incorporated. The four images, from left to right, corresponded to the undersampled image input (zero-filled and reconstructed using conventional inverse Fourier Transform), the fully sampled image (ground truth), the

images reconstructed using the convolutional neural network, and the image reconstructed using the self-attention convolutional neural network, respectively. The losses of micro-structures caused by sparse sampling, as appeared in the zero-filled undersampled image, were substantially recovered in both images reconstructed using deep learning approaches. Moreover, incorporation of the self-attention mechanism further improved the quality of reconstructed images, which was highly consistent with the ground truth. In subsequent experiments, the self-attention mechanism was always incorporated.

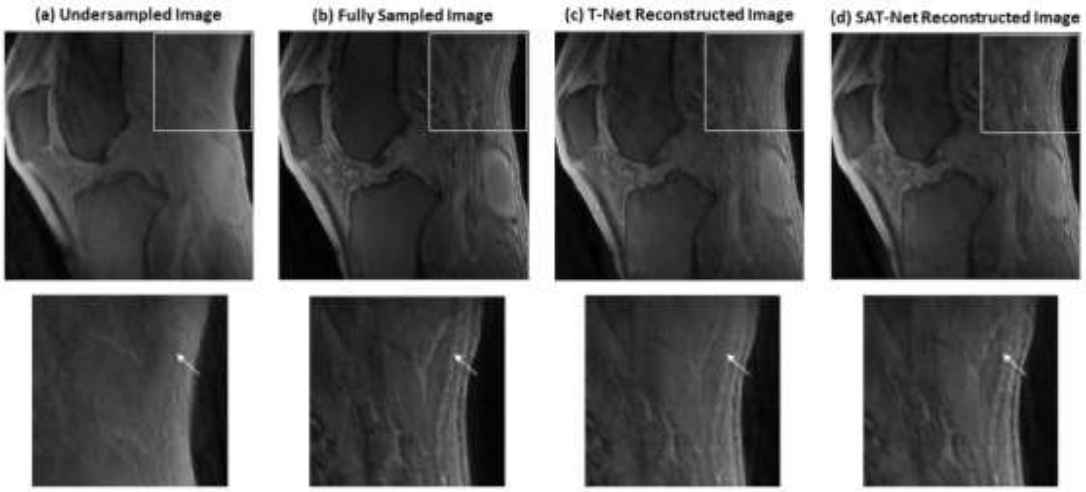



**Figure 8**. Comparison of images reconstructed with or without the self-attention mechanism incorporated, which were undersampled in a pseudo-random Cartesian sampling pattern with an acceleration factor of 6 achieved. (a) The undersampled image reconstructed using conventional Fourier Transform, which was the input to the system, (b) the fully sampled image, which was the ground truth, (c) the image reconstructed using the convolutional neural network T-Net, which was an impressive output, and (d) the image reconstructed using the self-attention convolutional neural network SAT-Net, which was a superior output. For each type of image, the whole picture was displayed in the upper row, and a region of interest was amplified in the lower row. The losses of micro-structures caused by sparsely sampling, as seen in the undersampled images, were substantially recovered in the images reconstructed using deep neural networks, as in (c) and (d). Incorporating the self-attention mechanism in (d) further improved the quality of the reconstructed image in (c), which had high fidelity with the ground truth (b).

In the training of neural networks, the Root Mean Squared Error (RMSE) was measured. The proposed SAT-Net had lower RMSE values than the T-Net, as shown in Figure 9.

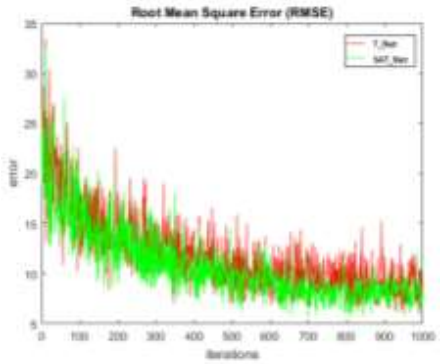



**Figure 9**. Comparison of Root Mean Squared Error (RMSE) in the training of the T-Net and the SAT-Net. The SAT-Net had lower RMSE than the T-Net.

The SSIM and PSNR of the 24 test images were shown in Figure 10. Images reconstructed using SAT-Net had higher SSIM and PSNR than those reconstructed using T-Net or the conventional zero-filling approach, demonstrating the improvement brought about by the incorporation of the self-attention mechanism.

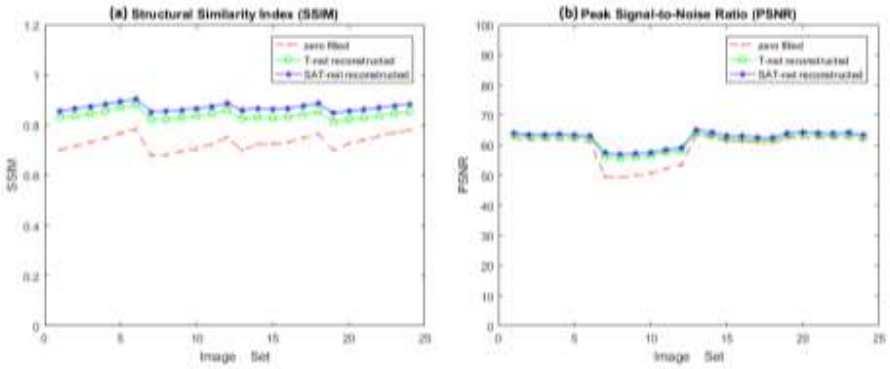



**Figure 10**. Quantitative performance evaluation of images reconstructed using different methods. (a) The SSIM of the test images reconstructed using zero-filling, T-Net, and SAT-Net, (b) the PSNR of the test images reconstructed using zero-filling, T-Net, and SAT-Net. Images reconstructed using SAT-Net had higher SSIM and PSNR than those reconstructed using T-Net or conventional zero-filling approaches.

The effect of using different patterns of local shortcut connections was investigated. Figure 11 showed images reconstructed using a self-attention convolutional neural network with no local shortcuts (as in U-Net), simple local shortcuts (as in V-Net), or relatively dense local shortcuts (as in T-Net and SAT-Net). The strategy of employing relatively dense local shortcuts did improve image quality over the other designs. The

same trend was confirmed quantitatively by measuring the average SSIM, PSNR, and RMSE, as shown in Table 1. The highest SSIM and PSNR, as well as the lowest RMSE, were achieved in images reconstructed with relatively dense local shortcuts (as employed in SAT-Net) compared to those obtained without local shortcuts (as in U-Net) or with simple shortcuts (as in V-Net).

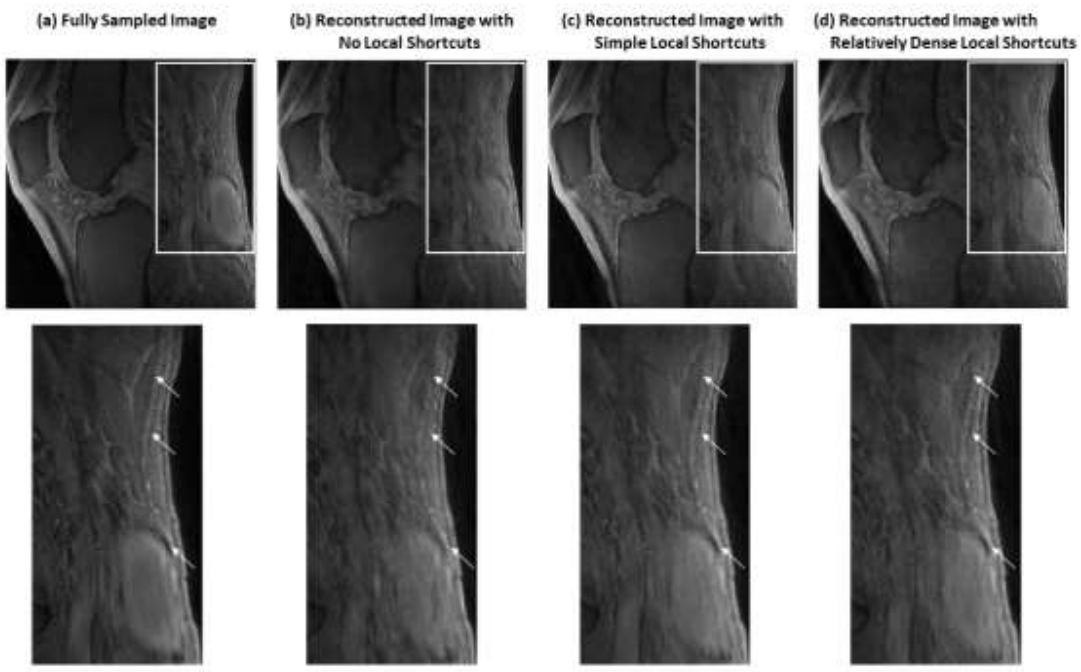



**Figure 11.** Comparison of images reconstructed with different local shortcut connections. (a) The fully sampled image, (b) the image reconstructed using a self-attention convolutional neural network without local shortcuts, as conducted in U-Net, (c) the image reconstructed using a self-attention convolutional neural network with simple local shortcuts, as conducted in V-Net, and (d) the image reconstructed using a self-attention convolutional neural network with relatively dense local shortcuts, as conducted in T-Net and SAT-Net. For each type of image, the whole picture was displayed in the upper row, and a region of interest was amplified in the lower row. The image reconstructed using a self-attention convolutional neural network with relatively dense shortcuts (d) was more similar to the ground truth (a) than the image reconstructed with a network that employed no local shortcuts (b) or simple local shortcuts (c).

| Images | SSIM | PSNR | RMSE |
|---|---|---|---|
| images reconstructed without local shortcuts | 0.82 | 60.22 | 9.83 |
| images reconstructed with single local shortcuts | 0.86 | 62.56 | 7.65 |

| | | | |
|---|---|---|---|
| images reconstructed with relatively dense local shortcuts | 0.87 | 63.23 | 6.82 |

**Table 1**. The average SSIM, PSNR, and RMSE of images reconstructed using self-attention deep convolutional neural networks with different shortcut connections. Images reconstructed with relatively dense local shortcut connections (as adopted in T-Net and SAT-Net) had higher SSIM and PSNR, as well as lower RMSE, than those reconstructed without local shortcuts (as in U-Net) or with simple local shortcuts (as in V-Net). This quantitative result was consistent with the observation in Figure 11.

In Figure 12, images reconstructed with or without data consistency enforced were compared to the ground truth. The enforcement of data consistency improved the quality of the reconstructed image.

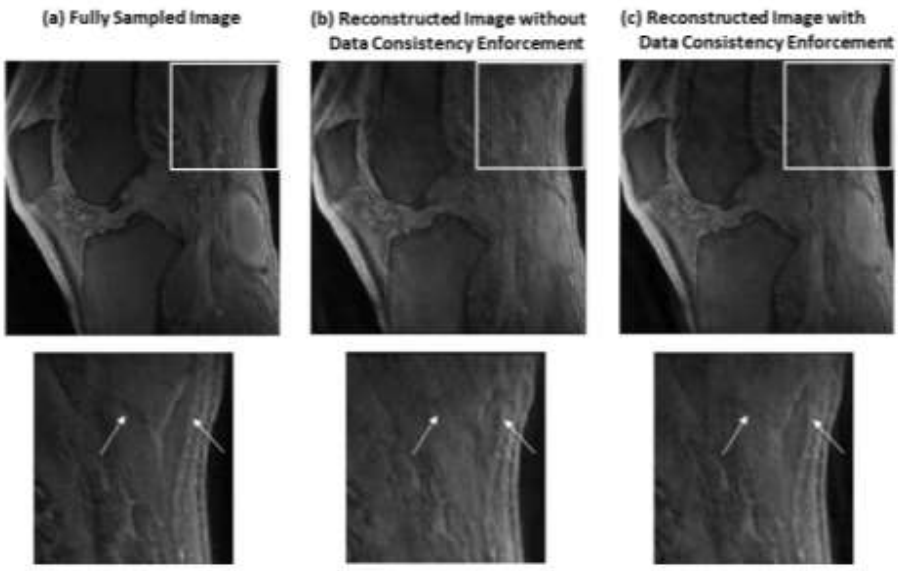



**Figure 12.** Comparison of images reconstructed with or without data consistency enforced against the ground truth. (a) The fully sampled image, (b) the image reconstructed using a self-attention convolutional neural network without data consistency enforced, and (c) the image reconstructed using a self-attention convolutional neural network with data consistency enforced. For each type of image, the whole picture was displayed in the upper row, and a region of interest was amplified in the lower row. The enforcement of data consistency clearly improved the quality of the reconstructed image.

## Discussion

In this study, we integrated the self-attention mechanism into a deep convolutional neural network for MRI reconstruction. It is advantageous to use the self-attention mechanism to capture long-range dependencies. In contrast to the progressive

propagation provided by a stack of convolutional operations, the self-attention mechanism computes direct interactions between any two positions, regardless of the distance. For this reason, the self-attention mechanism is computationally efficient.

When combined with deep learning, the self-attention mechanism could be built on top of a variety of deep neural networks for image processing applications. In the Image Transformer model (Parmar, 2018), one or more self-attention layers were combined with a fully connected layer to form the building block of an encoder-decoder network. However, the design of the fully connected layer did not suit image processing tasks since the large number of pixels caused heavy computation burden. To circumvent the problem, 'local' self-attention was performed on small image patches, which unfortunately had the effect of diminishing the strength of the self-attention mechanism as a means to model long-range dependencies. Alternatively, the self-attention mechanism was integrated into a conditional GAN to form the self-attention GAN (Zhang, 2018). While GANs (Goodfellow, 2014) have been extensively investigated for image synthesis, the training of GANs is well known to be unstable (despite numerous attempts to make their training more robust) and sensitive to the choices of hyper-parameters. In this work, we demonstrated the feasibility of integrating the self-attention mechanism into a convolutional neural network by appending a self-attention layer to every convolutional layer.

Certain design aspects of the T-Net model made it a particularly favorable convolutional neural network for MR image reconstruction. The first feature of T-Net was the relatively dense local shortcut connections constructed to facilitate residual learning. While the dense shortcuts in Dense Net boosted network performance by forwarding the output of every convolutional block to all subsequent blocks (Huang, 2017), it was computationally prohibitive for high resolution volumetric images. Inspired by Deep Recursive Residual Network (Tai, 2017), we developed our SAT-Net model which only forwarded the input of a hierarchical level to all the subsequent convolutional blocks at the same hierarchical level, improving the prediction outcome with an affordable computational cost. This specialized shortcut connection design helped SAT-Net achieve

an improved performance compared to the standard designs of U-Net and V-Net. The second aspect of T-Net which made it particularly well-suited for MR image resolution was that enforcement of data consistency was performed in k-space. Replacing predicted k-space coefficients with original measurements apparently improved the prediction accuracy. The efficacy of enforcing data consistency was demonstrated in the conjugated gradient highly constrained backprojection approach (Griswold, 2007). In the framework of deep learning, we enforced data consistency in a similar way to the work conducted in the cascade of convolutional neural networks (Schlemper, 2018).

In this study, the SAT-Net was trained with MR images acquired using a consistent imaging protocol. Compared with natural images or CT images, MR images that were consistently acquired were limited in number. We collected 336 three dimensional images and performed data augmentation to build a training set of a reasonable size. For each fully sampled image, different acceleration factors can be adopted, with higher acceleration factors expected to result in a decrease in image quality. In this study, we demonstrated an improvement in image fidelity at an acceleration factor of 6, which indicated a possibility to achieve a higher degree of acceleration.

The current approach can be extended in several directions. First, the self-attention convolutional neural network can be used to provide a direct mapping from k-space data to fully sampled images. In this study, we used zero-filled images as the input, which had undersampling artifacts. Secondly, alternative loss functions can be adopted in lieu of the RMSE that we used for this study. Potential candidates for loss function include $l_1$ loss, $l_2$ loss, SSIM, mutual information, or their combination. Thirdly, the self-attention convolutional neural network can be employed for specific MRI reconstruction tasks, such as dynamic MRI or quantitative MRI.

The proposed SAT-Net is a generic reconstruction framework that can be applied to various acquisition techniques (data sampling trajectories, pulse sequences, etc.) for diverse clinical applications. Moreover, the proposed self-attention convolutional neural network could potentially benefit tremendous image processing applications beyond MRI reconstruction due to the wide application of convolutional neural networks.

## Conclusion

A self-attention convolutional neural network framework was developed for the reconstruction of sparsely sampled MRI, aiming to provide improved image fidelity for accelerated MR image acquisition. Incorporation of the self-attention mechanism into convolutional neural networks effectively improved the reconstruction outcome by taking advantage of long-range dependencies. With additional support from relatively dense shortcut connections and data consistency enforcement, a high acceleration factor of 6 was achieved in cartilage MRI, while maintaining high image quality. The self-attention convolutional neural network architecture not only provides a generic framework for MRI reconstruction, but also has the potential to benefit other applications that employ convolutional neural networks.

## Acknowledgements

The authors would like to thank Dr. Cheng Tang for the helpful discussion. The research was supported by NIH/NCI (1R01 CA176553), NIH/NIAMS (1R01 AR068987), NIH/NINDS (1R01 NS092650), Varian Medical Systems, and Faculty Research Award from Google Inc.

## Reference

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265-283).

Bilgic, B., Goyal, V. K., & Adalsteinsson, E. (2011). Multi-contrast reconstruction with Bayesian compressed sensing. *Magnetic resonance in medicine*, *66*(6), 1601-1615.

Chaudhari, A. S., Fang, Z., Kogan, F., Wood, J., Stevens, K. J., Gibbons, E. K., ... & Hargreaves, B. A. (2018). Super-resolution musculoskeletal MRI using deep learning. *Magnetic resonance in medicine*, *80*(5), 2139-2154.

de Jesús Rubio, J. (2009). SOFMLS: online self-organizing fuzzy modified least-squares network. *IEEE Transactions on Fuzzy Systems*, *17*(6), 1296-1309.

de Jesús Rubio, J. (2017). Interpolation neural network model of a manufactured wind turbine. *Neural Computing and Applications*, *28*(8).

de Jesús Rubio, J., Lughofer, E., Meda-Campaña, J. A., Páramo, L. A., Novoa, J. F., & Pacheco, J. (2018). Neural network updating via argument Kalman filter for modeling of Takagi-Sugeno fuzzy models. *Journal of Intelligent & Fuzzy Systems*, (Preprint), 1-12.

Gao, X., Mu, T., Goulermas, J. Y., & Wang, M. (2018). Attention driven multi-modal similarity learning. *Information Sciences*, *432*, 530-542.

Gibson, E., Li, W., Sudre, C., Fidon, L., Shakir, D. I., Wang, G., ... & Whyntie, T. (2018). NiftyNet: a deep-learning platform for medical imaging. *Computer methods and programs in biomedicine*, *158*, 113-122.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

Griswold, M., Barkauskas, K., Blaimer, M., Moriguchi, H., Sunshine, J., & Duerk, J. (2007). More optimal HYPR reconstructions using a combination of HYPR and conjugate-gradient minimization. In *Proceedings of the International Society of Magnetic Resonance in Medicine* (pp. 19-25).

Hammernik, K., Klatzer, T., Kobler, E., Recht, M. P., Sodickson, D. K., Pock, T., & Knoll, F. (2018). Learning a variational network for reconstruction of accelerated MRI data. *Magnetic resonance in medicine*, *79*(6), 3055-3071.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700-4708).

Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), 61-70.

Ke, Z., Wang, S., Cheng, H., Ying, L., Liu, Q., Zheng, H., & Liang, D. (2019). CRDN: Cascaded Residual Dense Networks for Dynamic MR Imaging with Edge-enhanced Loss Constraint. *arXiv preprint arXiv:1901.06111*.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

Li, Y., Wang, X., Liu, W., & Feng, B. (2018). Deep attention network for joint hand gesture localization and recognition using static RGB-D images. *Information Sciences*, *441*, 66-78.

Liu, J., & Saloner, D. (2014). Accelerated MRI with CIRcular Cartesian UnderSampling (CIRCUS): a variable density Cartesian sampling strategy for compressed sensing and parallel imaging. *Quantitative imaging in medicine and surgery*, *4*(1), 57.

Liu, H., Xu, J., Wu, Y., Guo, Q., Ibragimov, B., & Xing, L. (2018). Learning deconvolutional deep neural network for high resolution medical image reconstruction. *Information Sciences*, *468*, 142-154.

Lodhi, B., & Kang, J. (2019). Multipath-DenseNet: A Supervised ensemble architecture of densely connected convolutional networks. *Information Sciences*, *482*, 63-72.

Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Lustig, M., Donoho, D., & Pauly, J. M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *58*(6), 1182-1195.

Lyu, Q., & Wang, G. (2018). Quantitative MRI: Absolute T1, T2 and Proton Density Parameters from Deep Learning. *arXiv preprint arXiv:1806.07453*.

Ma, Y. J., Carl, M., Searleman, A., Lu, X., Chang, E. Y., & Du, J. (2018). 3D adiabatic T1ρ prepared ultrashort echo time cones sequence for whole knee imaging. *Magnetic resonance in medicine*, *80*(4), 1429-1439.

Meng, X. L., Shi, F. G., & Yao, J. C. (2018). An inequality approach for evaluating decision making units with a fuzzy output. *Journal of Intelligent & Fuzzy Systems*, *34*(1), 459-465.

Milletari, F., Navab, N., & Ahmadi, S. A. (2016, October). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* (pp. 565-571). IEEE.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. *arXiv preprint arXiv:1802.05751*.

Parikh, A. P., Täckström, O., Das, D., & Uszkoreit, J. (2016). A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

Peters, D. C., Korosec, F. R., Grist, T. M., Block, W. F., Holden, J. E., Vigen, K. K., & Mistretta, C. A. (2000). Undersampled projection reconstruction applied to MR angiography. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, *43*(1), 91-101.

Pike, G. B., Meyer, C. H., Brosnan, T. J., & Pelc, N. J. (1994). Magnetic resonance velocity imaging using a fast spiral phase contrast sequence. *Magnetic resonance in medicine*, *32*(4), 476-483.

Qin, W., Wu, J., Han, F., Yuan, Y., Zhao, W., Ibragimov, B., ... & Xing, L. (2018). Superpixel-based and boundary-sensitive convolutional neural network for automated liver segmentation. *Physics in Medicine & Biology*, *63*(9), 095017.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Springer, Cham.

Schlemper, J., Caballero, J., Hajnal, J. V., Price, A. N., & Rueckert, D. (2018). A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE transactions on Medical Imaging*, *37*(2), 491-503.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Tai, Y., Yang, J., & Liu, X. (2017). Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition* (pp. 3147-3155).

Vaswani, N., & Lu, W. (2010). Modified-CS: Modifying compressive sensing for problems with partially known support. *IEEE Transactions on Signal Processing*, *58*(9), 4595-4607.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*(pp. 5998-6008).

Wang, S., Su, Z., Ying, L., Peng, X., Zhu, S., Liang, F., ... & Liang, D. (2016, April). Accelerating magnetic resonance imaging via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (pp. 514-517). IEEE.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794-7803). Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2018). Self-attention generative adversarial networks. arXiv preprint arXiv:1805.08318.

Xing, L., Krupinski, E. A., & Cai, J. (2018). Artificial intelligence will soon change the landscape of medical physics research and practice. *Medical physics*, *45*(5), 1791-1793.

Yang, Y., Sun, J., Li, H., & Xu, Z. (2017). ADMM-Net: A deep learning approach for compressive sensing MRI. *arXiv preprint arXiv:1705.06869*.

Yang, G., Yu, S., Dong, H., Slabaugh, G., Dragotti, P. L., Ye, X., ... & Firmin, D. (2018). DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE transactions on medical imaging*, *37*(6), 1310-1321.

Zhang, X. M., & Han, Q. L. (2018). State estimation for static neural networks with time-varying delays based on an improved reciprocally convex inequality. *IEEE transactions on neural networks and learning systems*, *29*(4), 1376-1381.

Zhu, B., Liu, J. Z., Cauley, S. F., Rosen, B. R., & Rosen, M. S. (2018). Image reconstruction by domain-transform manifold learning. *Nature*, *555*(7697), 487.